NÅGRA SPRÅKDATA UR EN PEDAGOGISK-PSYKOLOGISK VERKBAS

Inger Bierschenk

Bierschenk, I. Några språkdata ur en pedagogisk-psykologisk verkbas.
Stencil (Malmö: Pedagogisk-psykologiska institutionen), December 1978.

Denna arbetsrapport från Informations- och dokumentationsprojektet (I&D) sammanfattar några studier som gjorts på ett sampel utbildningsforskares 40-åriga produktion.
Studierna omfattar kvantitativa fördelningar av de språk som skrifterna är avfattade på och redovisar resultat från ett program med automatisk språkbestämning. Vidare jämförs de olika språkens separeringsförmåga i förhållande till ett vetenskapligt verks bibliografiska enheter. Ordfrekvenser inom dessa s k dataposter diskuteras med avseende på innehåll respektive struktur och några jämförelser med andra större textmaterial avslutar rapporten.

Nyckelord: Datalingvistik, datorbaserad lexikologi, forskningsinformation, jämförande lingvistik, pedagogisk dokumentation

INNEHÅLL

1. BESTÄMNING AV SPRÅKET I VETENSKAPLIGA VERKBESKRIVNINGAR

Projektet "Information och dokumentation" (I&D) syftar bl a till att utveckla program för automatiska hanteranden av dokumentbeskrivningar till forskningsrapporter inom utbildningssektorn. Dokumenten är de skrifter som ett sampel forskare producerat under en 40-årsperiod. I&D-projektet anknyter till ett tidigare projekt, där urval och övriga bakgrundsdata har beskrivits (B. Bierschenk, 1974). En utförlig beskrivning av I&D-projektets databaser finns i B. Bierschenk (1978). Baserna består av 949 verk och i dem förekommande referenser (ca 23 000).

Beskrivningarna till dessa verk utgörs av bibliografiska data, dvs sådana uppgifter som refereringskonventionerna bjuder angående författarnamnet, titeln, utgivningsort, förlag, årtal, m m (se B. Bierschenk, 1978, kodningsanvisningar i bil 1). Dessa data har delats upp i s k poster, så att viss specifikation alltid förekommer tillsammans med viss bestämd sifferkod. På vilket språk ett verk är avfattat avgörs bäst genom att man läser och identifierar språket i titeln. Varken författarnamn, tryckort, förlag eller liknande bibliografiska data är särskilt säkra uppgifter för detta ändamål. Angående betydelsen av dessa tre i förhållande till övriga hänvisas till I. Bierschenk (1978). Innan vi går in på detaljer i en sådan språkbestämning ger vi i tabell 1 fördelningen av olika språk i svensk utbildningsforskning.

Tabell 1. Fördelning av verktitlar på olika språk

	Sv	Eng	Ty	Fr	No	Da	Övr	Σ
Antal	660	249	26	5	2	4	3	949
%	70	26	3	1	-	-	-	100

Tabell 1 visar att svenska språket dominerar starkt när de 40 forskarna skriver sina forskningsrapporter och böcker. En fjärdedel är engelska. Endast några få forskare producerar sig på tyska. De övriga språken står i stort sett två forskare för. I Bierschenk (1978) kunde bl a konstatera att forskarna refererar engelsk litteratur till 50 % och svensk till 28 %.

Övriga 22 % är referenser på 11 språk. Så stor variation blir det inte i de egna verken, ett resultat som väl inte är förvånande. Bortsett från att det kan vara en kunskaps- eller färdighetsfråga, så är det ju också kostnader förknippade med översättningsarbeten.

Inom projektet har utarbetats en algoritm för en automatisk separering av referenser i olika språk (se I. Bierschenk, 1978). Programmet har testats i olika omgångar på såväl referenser som verkbas på det material som avsåg 1937-1974, men har endast presenterats för referensbasen under denna tidsperiod. En kompletterande insamling fullföljdes under sommaren 1978. I den ovan anförda rapporten gjordes kontroller av det nytillkomna materialet i syfte att spåra skillnader jämfört med det stora materialet. Några kontroller tydde bl a på att språkbestämningsprogrammet verkade något bättre på verk än på deras referenser, troligen till följd av att variationerna i språkligt hänseende inte är så stora i det mindre materialet.

Resultat för verkbasen presenterades inte i den tidigare gjorda studien, varför denna arbetsrapport ska tala om vilka språkliga data av detta speciella slag som finns i de 40 forskarnas egen produktion. Presentationen inbegriper dessutom åren 1975-77.

För att bestämma språket i ett verk har ett antal "sökord" (teckensträng) använts. Hur dessa tillkommit redovisas i I. Bierschenk (1978, ruta 8, s 20) och upprepas inte här. Däremot ska det sägas, att söktekniken innebär att en strängmatchning från vänster sker. Det först påträffade ordet (trunkerat för ändelser, etc) som är likadant som söksträngen fäller utslaget, dvs orsakar att hela verket tilldelas en språkkod. Samtidigt räknas frekvenser för de matchande sökorden. Det betyder alltså att andra sökord kan finnas i samma verk utan att de får betydelse för den automatiska sorteringen.

Detta språkbestämningsprogram har nu testats på hela verkbasen (n = 949), som jämfört med den första rapporten innefattar tidsperioden 1937-1977, dvs en 40-årsperiod inom utbildningsforskningen.

Vissa verk blir via programmet inte bestämda, eftersom vi inte ännu har utvecklat andra sökrutiner än enkla matchningar. Kategorin "obestämda" betyder således inte felaktiga. Resultatet visas i tabell 2.

Tabell 2. Resultat av språkbestämning av verkbasen

	Sv		Eng		Ty		Fr		No		Da		Obest		Σ	
	f	%	f	%	f	%	f	%	f	%	f	%	f	%	f	%
Korrekta utfall	572	67	244	29	25	3	3	-	1	-	1	-			846	89
Felaktiga utfall	1		1		1		0		0		0				3	0
Obest													100	11		11
Σ	573	60	245	26	26	3	3	-	1	-	1	-	100	11	949	100

Tabell 2 visar att 89 % av verken blir korrekt bestämda. De tre felen är obetydliga: Trunkeringen ein* har givit ett norskt verk vars titel innehåller namnet Einar. Felet i den svenska listan är en dansk titel, som kommer ut genom sökordet *pedagogik* . Det "engelska" felet är en svensk titel med fyra engelska ord i början. Dessa fel är kända sedan de tidigare utprövningarna. Tillsammans svarar svenska och engelska verk för 86 % av utfallet. För att ge en uppfattning om vad den obestämda listan innehåller visar tabell 3 en fördelning av de 100 obestämda.

Tabell 3. Fördelningen av olika språk i de obestämda verken

	Sv	Eng	Fr	Övr västspr	Da	Ty	Övr östspr	No	La	Fi	Σ
Antal verk	87	5	2	2	2	1	1	0	0	0	100

Dessa olika språkkoder har använts på den större basen, som bl a innehåller referenser på latin och finska. Bland "övriga västspråk" finner vi i verken italienska (i referensbasen även spanska och holländska). De "övriga östspråken" är ryska och tjeckiska.

Som tabell 3 visar, kommer en utvidgning av programmet främst att gälla regler för svenska verk, som alltså utgör 87 % av dem som programmet ännu inte tar hand om.

2. DATAPOSTERNAS BETYDELSE

Den utvärdering av programmet för automatisk språkseparering som gjorts på referensbasen har redovisat de olika posternas betydelse för språkseparering. "Post" betyder en bibliografisk enhet, uppdelad enligt sitt innehåll, t ex "titel" och "tryckort" (se I. Bierschenk, 1978, ruta 1). Även om våra frekvenser är få, borde vi ändå kunna se om utfallet för verken ser likadant ut som för referenserna. Tabell 4 får därför presentera datapostemas betydelse (posterna 1 = författare, 4 = förlag och 5 = årtal har ej använts).

Tabell 4. Dataposternas betydelse vid språkseparering av verk

	Antal titlar inom post							
	2		3		6		Σ	
	f	%	f	%	f	%	f	%
Svenska	555	66	-		16	2	571	67
Norska	-		1	0	-		1	0
Danska	-		-		1		1	0
Tyska	23	3	-		2		25	3
Engelska	244	29	-		-		244	29
Franska	3		-		-		3	0
Σ	825	98	1	0	19	2	845	100

Fördelningen avser korrekta utfall (jfr tab 2)
Post 2 = titel, 3 = tryckort, 6 = tidskrift/institution

Titelposten tar upp 98 % av verken (för referenserna var siffran 96 %). Liksom referenserna tas norska och danska ut via andra poster än titelposten. (Men det är givetvis svårt att jämföra eftersom vi här endast har en frekvens vardera för norska och danska.) Tendensen att svenska dokument även i rätt hög grad tas upp från post 6 håller för både referenserna och verken. De engelska finns enbart inom titeln, dvs engelska språket är ännu mer koncentrerat än vad det såg ut för referenserna. Det sprider inte ut sig på fler poster än en, vilket är naturligt, eftersom de svenska verken på engelska inte utges i t ex New York (post 3 har inte använts).

Vi ser alltså att reglerna i första hand kan koncentrera sig på "cues" (sökord) inom titeln, dvs de språkspecifika cues. Referensspecifika ("cues" inom poster som hänför sig till refereringskonventionerna) är bara av intresse för svenska institutionsrapporter.

3. INNEHÅLLSORD ELLER FORMORD SOM SPRÅKBESTÄMMARE

I den refererade studien över automatisk språkbestämning av referenser diskuterades bl a typen av sökord för bästa utfall på de olika språken. Det kunde konstateras att orden tillhörde minst två kategorier, nämligen formord, som anger språkstrukturella egenskaper i titeln, och ämnes- eller innehållsord, som anger vad ett dokument avhandlar i förhållande till ämnesstrukturen. Söktekniken visade att vissa formord ger det bästa resultatet, dvs sådana ord som även i annat språkligt material har hög frekvens. Därefter kommer sådana innehållsliga ord som inom utbildningsforskningen betraktas som självklara, t ex "pedagogik". I det följande ges listor över sökord som haft betydelse inom de referens-specifika posterna (3 och 6) respektive den språkspecifika titelposten.

Tabell 5. Frekvenslista för sökord på posterna 3 och 6 (samtliga språk)

Sökord	Post	f	%
Oslo	3	1	5
dansk	6	1	5
Z	6	1	5
und	6	1	5
lärar*	6	7	35
och	6	3	15
SOU	6	2	10
års*	6	2	10
svensk	6	2	10
Σ		20	100

Det sökord som utmärker sig tydligast är ämnesordet lärar*. En kontroll visar att det kommer från institutionsangivelser, t ex "lärarhögskolan i Malmö", och innebär att dokumentet är en rapport eller stencil, utgiven eller avfattad vid en institution. Formordet och har också en viss betydelse, beroende på dess vanlighet i svenska språket överhuvudtaget (jfr I. Bierschenk, 1978). Förekomsten inom post 6 innebär t ex "Tidskrift för psykologi och pedagogik".

I referensbasen kunde en rangordning märkas mellan dessa svenska sökord. lärar* har mycket större betydelse i verken än i referenserna.

Det verkar helt naturligt. *svensk* har i referenserna större betydelse (tar upp de flesta). I verken utmärker sig inte *svensk* framför något annat. lärar* har intagit den platsen. Dessa fem svenska sökord är dock fortfarande de fem som har någon betydelse för svenskspråkiga dokument inom utbildningsforskningen, när det gäller att fånga upp dem utanför vad titeln kan ge. Men titeln ger trots allt 97 % av de svenska verken, varför en koncentration till de språkspecifika titelorden bör ske.

Tabell 6. Frekvenslista för sökord på post 2 (svenska)

Sökord	f	%	Sökord	f	%
och	126	22	vux*	8	1
skola	59	10	års*	8	1
för	52	9	barn*	8	1
till	34	6	begåvning*	7	1
psykologi	31	5	*utredning*	7	1
*utbildning	22	4	hos	6	1
ett	21	4	läro*	6	1
pedagogik	19	3	*pedagogiska*	5	1
mätning	16	3	eller	4	1
någ*	16	3	är	3	1
att	15	3	samt	3	1
*undersökning	13	2	uppfostr*	3	1
arbet	12	2	personlighet*	3	1
svensk	11	2	ur	2	0
vad	10	2	Sverige	2	0
inom	10	2	hur	1	0
mellan	10	2	från	1	0
			inför	1	0

Totalsumma korrekta svenska = 571
(-1 för *pedagogik*)

Här ser vi att och har lika stor andel här som i referenserna. Men medan *svensk* i referenserna tog 7 % (rangplats 3), så tar ordet i verken endast 2 % (rangplats 14). Barn* har ändrat sig från referensernas 4 % till verkens 1 %. Orden beteende, betänkande och historia finns inte med. Huruvida de ändå finns i materialet kan denna frekvensräkning inte svara på eftersom söktekniken från vänster antecknar en frekvens på det först påträffade sökordet. Därför ska vi inte heller spekulera vidare i begreppens eventuella förändringar.

Tabell 7 visar frekvenserna för engelska sökord.

Tabell 7. Frekvenslista för sökord på post 2 (engelska)

Sökord	f	%	Sökord	f	%
the	47	19	some	6	2
of	47	19	studies	5	2
and	35	14	measur*	5	2
educational	14	6	learning	4	2
study	14	6	from	3	1
on	12	5	methods	2	1
school*	12	5	psychological	2	1
analysis	9	4	reading	2	1
education	9	4	by	1	0
research	6	2	or	1	0
teach*	6	2	with	1	0
			child	1	0

Totalsumma korrekta engelska = 244
(-1 för the)

Sökorden the, of och and hade även i referensbasen största delen av utfallet, men skillnaderna mellan dem och övriga ord var större i den stora basen. I verkbasen ligger the och of på samma frekvens, medan the i referenserna hade nio procentenheter fler än of.

Vad vi i övrigt kan lägga märke till är att några ämnesord inte funnits med (i detta sökprogram) t ex personality, psychology, adult och training. Men dessa hade obetydlig frekvens även i referenserna.

Vi har i verkbasen inte många titlar på tyska, och ännu färre på franska. De franska söksträngarna är l', som har tagit två titlar och les, som tagit ett (totalt tre). De tyska som har givit utslag redovisas i tabell 8.

Tabell 8. Frekvenslista för sökord på post 2 (tyska)

Sökord	f	%	Sökord	f	%
schul*	6	26	bildung	2	9
und	5	22	über	2	9
die	3	13	das	1	4
forschung	3	13	*buch*	1	4
			(ein*	1)	

Totalsumma korrekta tyska = 23
(-1 för ein*)

De tyska und och die kan jämföras med engelskans and och the. schul* , school* och *skola* finns med i toppen för samtliga betydelsefulla språk. De flesta tyska sökord har inte fått några frekvenser, beroende på för litet antal verk. Anmärkningsvärt är bl a att *forschung* i verken står för 13 % men i referenserna för 0. Det visar kanske att översättningar (som det måste röra sig om) slår ut här, dvs begreppen är bildade för svenska förhållanden.

Sammanfattningsvis kan vi konstatera att några få formord toppar listorna, avbrutna av mycket övergripande (och ur informationssynpunkt ointressanta) ämnesord. Övriga formord finns insprängda längre ner i listorna och har låg frekvens enligt vår speciella sökteknik.

Vid kontrollerna av språksepareringsprogrammet gjordes en översiktskontroll av hur formorden respektive ämnesorden förhöll sig i ett tidsperspektiv (I. Bierschenk, 1978, kap 4. 3). Några ämnesord förändrar sig såtillvida att de byter rang som separatorer. Formorden förändras däremot inte nämnvärt. Slutsatsen som gjordes var att strukturen i titlar tycks vara konstant över en lång tidsperiod, vilket gör att studier av begreppsförändringar blir meningsfulla. Sådana studier är påbörjade. Dessutom har vi ansett oss behöva studera själva strukturorden, främst konjunktioner och prepositioner, dels på grund av deras höga frekvens allmänt, vilket torde vara användbart vid automatiska sorteringar av olika slag, dels på grund av deras konstanta uppträdande i titlar till forskningsrapporter. En särskild studie kommer därför att ägnas dessa typer av formord. Några förutsättningar diskuteras i det sista kapitlet.

4. EXPLORERANDE JÄMFÖRELSER MELLAN FREKVENSSTUDIER

Såsom det förra kapitlet redovisade har vissa formord i titlar hög separeringsförmåga, på grund av hög frekvens. Det finns dessutom många andra som skulle kunna användas som "nycklar" för olika sorteringsuppgifter, särskilt nu när vi inte behöver ta hänsyn till tvetydigheten (separeringen är ju färdig). Mycket talar för att de strukturella orden uppträder ganska stereotypt i titlar. Somliga förekommer i början, andra har hellre sin plats längre bak, kanske beroende på om något annat formord finns där eller ej. Detta skulle kunna gälla för prepositioner. Konjunktionerna sammanbinder och borde uppträda i mitten, nämligen mellan de begrepp de sammanbinder.

Vad som intresserar är således strukturorden, inte vilka formord som helst. (Vi bortser från artiklar, etc.) Prepositioners och konjunktioners frekvenser i verktitlarna har tagits ut och fördelar sig som tabell 9 visar.

Tabell 9. Strukturords frekvenser i titlarna

Ord	f	%	Ord	f	%
och	296	25	rörande	12	0
i	222	18	eller	8	0
av	178	15	samt	7	0
för	82	7	kring	6	0
på	65	6	bland	5	0
till	58	5	ur	4	0
med	53	4	efter	3	0
om	45	4	mot	3	0
inom	32	3	genom	3	0
vid	28	2	enligt	3	0
som	25	2	inför	2	0
mellan	22	2	över	1	0
från	17	1			
hos	13	1			

Totala antalet strukturord = 1 203

Vi ser här att tre strukturord dominerar, att det finns en mellangrupp samt att ungefär hälften har mycket låg andel i titlarna. För olika strukturanalyser har en konkordans utförts på samtliga dessa ord, där det bl a visar sig att de första åtta orden tycks finnas med i de mest

typiska titlarna.

För att spegla deras uppträdande i andra skrivna texter har en sammanfattning gjorts av rangordningen mellan de åtta orden som utläses ur tre frekvensordböcker för svenska språket (Widegren, 1935; Hassler-Göransson, 1966; Allén, 1970) och jämfört med förekomsten i våra verktitlar. Jämförelsen är intressant därför att frekvenserna i de fyra materialen är uppbyggda med olika förutsättningar: Widegren och Allén redovisar t ex homograferna separerade. Andra skillnader är bastalet och de olika texttyperna (debattspråk, brev, uppsatser, skönlitteratur och tidningstext) samt de samplade åren (från 1920-talet till 1970-talet). Vårt material har dessutom inga uppgifter om ordantal, även om det går att skatta. Vårt bastal är summan av de åtta ordens förekomst i 949 verk. I samtliga listor utom Hassler-Göransson har i förekommande fall uppgifterna för prepositioner tagits (således inte uppgift som gäller samma ord t ex i funktionen adverb). Frekvenserna resulterar i följande rangordning mellan strukturorden.

Tabell 10. Ordfrekvenser som bas för rangordning av strukturord i svenskt skriftspråk

			Rang	
Ord	W	H-G	A	I&D
av	3	7	3	3
för	4	5	5	4
i	1	2	2	2
med	7	6	6	7
och	2	1	1	1
om	8	8	8	8
på	5	3	4	5
till	6	4	7	6

W = Widegren, H-G = Hassler-Göransson,
A = Allén, I&D = I&D-projektets forskningsmaterial

Vi kan läsa tabellen på flera sätt, men för vårt syfte räcker det med att konstatera att likheterna är störst mellan forskningsspråk i titlar (1937-1977) och debattspråket i riksdagen (1920-30) med avseende på strukturord. Ju mer skönlitterär en text är desto färre blir likheterna. Tidningsspråket år 1965 (Allén) innehåller också i sig så pass blandade textavsnitt att en jämförelse egentligen är svår. Trots detta finns det

stora likheter. Titlar på forskningsrapporter är i hög grad formaliserade, vilket gör att likheten med Hassler-Göranssons lista inte blir så hög även om det finns ett påvisbart samband, trots att hon inte har separerat homografer.

De ord som uppvisar "osäkerhet" är av, på och till sett över alla fyra materialen. En gissning är att dessa tre har mer kontextuell variation än övriga. Jämför vi de två mest lika materialen är det bara en rangskiftning mellan i och och.

Utan att gå in på statistiska spörsmål kan sägas att likheten i textmängden mellan W och H-G är störst, liksom årtalet för samplingen, men sambandet mellan orden är lägst. Det tycks alltså som om likheten i texttyp har större betydelse vid en jämförelse av strukturord än texternas ålder.

För vårt vidkommande betyder detta att strukturorden får stor betydelse vid utveckling av en generell modell för informationsåtervinning, där materialet speglar en 40-årig utveckling. Dessutom skulle denna modell vara tillämpbar på annat material av samma typ, t ex forskningsrapporter i andra ämnen.

För att få en mera tillförlitlig förankring av siffrorna i tabell 10 presenteras här en rangkorrelation med parvisa jämförelser samt korrelationen mellan samtliga fyra ordlistor.

Tabell 11. Rangkorrelation på fyra textmaterials rang av strukturord

Jämförelse-par		Rang-korrelation	z
W— H-G	r_s	.67	1.76
W— A	r_s	.93	2.46
W— I&D	r_s	.98	2.58
H— G—A	r_s	.69	1.83
H-G—I&D	r_s	.69	1.83
A—I&D	r_s	.95	2.52
W—H-G—A—I&D	W	.86	χ^2 24.17

Korrelationerna återspeglar det förhållande som diskussionen ovan gav, men de stärker oss ytterligare i våra antaganden för en fortsatt bearbetning. Dessa korrelationer säger nämligen att sambandet mellan de fyra materialen är mycket högt, nästan perfekt mellan I&D och W respektive A. H-G avviker från de övriga tre, men totalt sett är strukturordens uppträdande tämligen konstant. I vilken mån de tre orden av, på och till har någon betydelse visar kanske fortsatta studier av konkordanserna, som kommer att ge närmare upplysningar om vilka strukturer som kännetecknar titlar till vetenskapliga verk inom utbildningsforskningen.

5. REFERENSER

Allén, S. et al. Nusvensk frekvensordbok baserad på tidningstext. Del 2. Stockholm: Almqvist & Wiksell, 1971.

Bierschenk, B. Perception, strukturering och precisering av pedagogiska och psykologiska forskningsproblem på pedagogiska institutioner i Sverige. Pedagogisk-psykologiska problem, Nr 254, 1974.

Bierschenk, B. En longitudinell analys av kunskapsutvecklingen inom utbildningsforskningen. Stencil (Malmö: Pedagogisk-psykologiska institutionen), December, 1978.

Bierschenk, I. Försök med automatisk separering av referenser i en flerspråkig databas. Testkonstruktion och testdata, Nr 34, 1978.

Hassler-Göransson, C. Ordfrekvenser i nusvenskt skriftspråk. Lund: Skriptor, 1966.

Widegren, P.G. Frekvenser i nusvenskans debattspråk. Stockholm: Kungl. Maj:T, m fl, 1935.